SEDAC Inputs to SEEDS Levels of Service Workshop

Robert S. Chen, SEDAC Manager
W. Christopher Lenhardt, Deputy SEDAC Manager

SEDAC is pleased to have the opportunity to provide inputs into the SEEDS Levels of Service Workshop and to comment on the SEEDS Requirements/LOS & Cost Model Working Paper Draft (dated 12/21/2001). We commend the SEEDS team for initially focusing on requirements issues and for seeking to balance overall system requirements with the essential role of data centers in the future evolution of Earth Science Enterprise data and information systems.

However, we are struck by the somewhat narrow framework and near-term perspective of the requirements analysis, which seems to be planning a system based on assumptions and *modis operandi* from the past decade when it is very clear that not only technology is changing rapidly, but also the underlying approach to science missions and the ways in which users access and utilize scientific data.

Changing Assumptions about Science Missions

EOS missions like Terra and Aqua were premised on the need for simultaneous observations from multiple instruments on a large platform; minimizing risk and providing high capacity throughput from the satellite to the ground-based archives were therefore dominant drivers in the design of EOSDIS.

Current ESE planning is concentrated on smaller, more focused missions, with instruments geared to specific science and applied needs. When simultaneous measurements are desirable, data integration may be performed based on high quality calibration and georeferencing (e.g., ESE, 2000a,b).

Similarly, current missions focus on producing science quality products in near-real time, e.g., MODIS Level 2 products are generally made available 24 hours or more after the observations are taken. However, many possible applications require more rapid turnaround, e.g., wildfire and flood monitoring; thus, there is currently a high level of interest in the MODIS direct broadcast capability around the world. Assembling a regional or global data product from individual ground stations with only a short delay (hours or even minutes) is now feasible using the Internet, drawing on relatively inexpensive receiving stations and processing capabilities.

There is currently much pressure to reduce the time before "validated" science products are made available for a wider range of scientific and applied users. Extrapolating this trend into the future, it seems likely that more attention will have to be given to rapid, automated validation and quality control and intercomparison with complementary data sources, e.g., *in situ* observations and data from operational satellites (both government and commercial).

How missions will evolve over the next two decades is not entirely clear. The development of "sensor webs" is one direction currently being explored by NASA and others; this could entail the need to manage data collected simultaneously by hundreds if not thousands of distributed

sensors with potentially disparate data streams. Such webs might also entail much greater two-way flows of data, e.g., in response to the need for real-time control of instrument parameters.

Current scientific interest in "regional" environmental variability and change and human interactions under discussion in the scientific community (e.g., NRC, 2001) could lead to more complex, regionally focused sets of measurements drawn from a variety of space-based, airborne, and ground-based instruments. Such research efforts could mark a return to the "campaign"-oriented data activities of the past, though with a more open-ended time frame. What seems likely is that a wider range of data sources, drawn from both the scientific and non-science communities and involving a wider range of scientific disciplines and expertise, will be of interest to both scientific and applied users.

These changes in mission assumptions are likely to have important implications for future data systems and data centers. For one, there is likely to be an increased emphasis on managing and integrating distributed, disparate data streams, rather than one central "pipe". Supporting data interoperability and allowing for increased access to data and its attendant documentation at various stages of processing will become important requirements. Maximum data flows may be harder to predict, suggesting the need for load balancing strategies that reduce capital costs and rely on distributed processing and network capacity, perhaps from commercial or other nongovernmental organizations. Increasing emphasis on cost recovery from a wider range of data users will generate stronger requirements for accountability, metrics, privacy, and interoperability with outside systems. With distributed data sources and systems comes increased needs for ensuring overall system reliability and integrity, especially in a world beset by hackers and terrorist threats; therefore, requirements for real-time authentication and security are likely to become increasingly important from both reliability and cost perspectives.

Changing Assumptions about User Needs and Expectations

User expectations about data and information access are clearly changing rapidly, both in the sciences and in the world at large. The rapid expansion of the Internet in terms of capacity, functionality, and content will continue to affect what data users expect to access, how they access them, and how rapidly they can utilize the data for productive activities. The continued implementation of the National Spatial Data Infrastructure, and more generally the Global Spatial Data Infrastructure, should make a wider range of data and information resources generally accessible to both scientists and non-scientists. Rapid reductions in the cost of data storage, processing, and visualization and analysis may continue to change user willingness to deal with large volumes of data and their perceived need for more real-time data and information.

In traditional Principal Investigator (PI) driven research, small teams of researchers consisting of one or two lead investigators and typically including a postdoctoral scientist, graduate students, and technical staff acquire copies of datasets of interest from multiple sources and integrate these into their own research environment, e.g., a set of computer hardware, "home-grown" modeling and analysis tools, and some selection of "off the shelf" software. Relatively few non-science groups had the capabilities to process or sensibly utilize the primary data, but might be able to utilize reprocessed data products coming from research groups.

Over the past decade, multi-PI research teams have become increasingly prevalent, consisting of a variety of investigators located around the U.S. and the world. In many cases, these teams have designated data managers and/or dedicated facilities for data processing, as well as direct relationships with disciplinary data centers. These teams are certainly capable of dealing with higher volumes of data on an ongoing basis, and have greater needs for real-time access to data processing streams for validation, real-time modeling, and, in some cases, forecasting.

Simultaneously, a wider range of users has emerged, capable of dealing with higher volumes of complex data, but not all driven by global change research interests. In some cases, these users utilize commercial software with fairly sophisticated functionality and growing capacity. Some of these are interested in obtaining large volumes of data, but many are also interested in more direct access to relevant subsets of data, e.g., for particular regions, time periods, or parameters or utilizing data mining techniques.[1] Many of these users have their own sources of spatial data, and remote sensing data are only one component of their activities and interests.

In the future, we believe that both scientific and applied users will want greater flexibility in data access. Traditional notions of well-defined and orderable datasets and data granules will become increasingly fuzzy. Data processing and analysis tools will increase in sophistication, capable of handling—and indeed requiring—more complex data inputs (e.g., with a greater range of ancillary data) and able to seamlessly integrate data with much higher levels of automation. Current efforts such as the Open GIS Consortium (OGC) Open Web Services initiative will permit users to identify and access data from multiple sources, pipe data automatically to distributed centers for processing according to user-specified algorithms, and deliver resulting outputs to a user's desktop or handheld computer, a computer model, or an intelligent client, accompanied by appropriate documentation and quality control information and supporting metrics and billing information. Predicting specific user capabilities 5 to 10 years in advance (or even 2-3 years!) is certainly difficult, but it seems likely that both scientists and non-scientists will have a wide array of powerful, flexible data discovery, access, analysis, visualization, and decision support tools that we can only partially envision today.

Users may also be much more willing and able to accommodate modest charges for data to support cost recovery, utilizing evolving modes of electronic commerce that can efficiently and securely track usage and transfer funds with little user intervention. This suggests that there will be opportunities to develop and implement a self-sustaining or partly self-sustaining business model in future system architectures.

What Probably Won't Change

Despite the above observations, we do believe that the need for a coordinated data and information system still exists, and that data centers are central to the evolution of this system.

The 1998 National Research Council review of NASA's Distributed Active Archive Centers emphasized the role of EOSDIS as a "tool for achieving the science goals of the Earth Science Enterprise" in facilitating the creation of data products, catalyzing the preparation of secondary

[1] For example, we are already aware of one national news organization that receives real-time seismic data that it is able to integrate with other spatial data and computer models in order to provide rapid, 3-dimensional graphics of seismic events and their potential impacts on people and infrastructure within hours of the event detection.

data sets and information products, making data readily available to the broader scientific community, and preserving data in usable form for future generations of scientists (NRC, 1998: 3). The report recognized the key role of the DAACs in meeting these objectives, serving both as "discipline centers that serve the needs of a relatively small, specialized constituency" and as "elements of a larger system, which serves the broader earth science community" (NRC, 1998: 28). Although the report noted weaknesses in collaboration and long-term strategic planning among the DAACs, it generally commended the DAACs for their strong linkages with the scientific community, responsiveness to users, and professional data management.[2]

We argue that the role of the data center as a focal point for hands-on expertise and long-term responsibility for data management remains vital to the continuing success of the evolving ESE data and information system. Even if users can find, access, and assimilate data using automated tools or intelligent agents, their successful use of the data will depend in large part on the underlying quality of the data management and associated documentation. With an increasing proliferation of data types, data sources, and value-added data processing tools, well-trained, accessible experts are needed to help users figure out appropriate data and analysis methods for their applications. This same proliferation of data sources and types could lead to confusion and duplication in data management and version control and potential loss of data that "fall between the cracks." Strong centers of expertise with hands-on experience with the range of data important to particular scientific disciplines or application areas will still be needed, with specific responsibility and resources for long-term data stewardship. Such centers would also have the background and perspective needed to allocate limited resources sensibly across competing needs, e.g., for data rescue, documentation, access, preservation, security, and user support, and across different user groups, e.g., scientists, applied users, and educational users. They should also have the long time perspective needed to promote data continuity and preservation appropriate to the needs of their science and user communities.

These centers should form the backbone for a flexible, adaptable data and information system that is responsive to both mission and user needs. The latter could be designed and constructed by a centralized group, but might also emerge through a more distributed, evolutionary process. Either way, the data centers are likely to fulfill persistent needs for data management expertise, scientific coordination, user support, technology infusion, and strategic planning.

Implications for Requirements and Levels of Service

This perspective suggests that the focus on requirements and levels of service should not be on the physical process of data center operations, which is where the current draft working paper concentrates, but on the higher-level functions needed to maintain the viability and effectiveness of the data centers and the overall system. Specifically, it is clear that all data centers need, in addition to general data center administration:

- scientific expertise in their areas of activity;
- data management expertise related to appropriate archiving, documentation, and distribution of data;

[2] Note: SEDAC was not included in this review due to its recompetition during the same time period.

- information technology expertise related to maintaining and supporting data and information systems; and
- user support expertise concerned with user access and support needs.

Requirements should be driven in part by recognized standards and policies, e.g., for data archiving, user access, external guidance, quality control, information technology security, data documentation, protection of intellectual property, and privacy and confidentiality. Emerging requirements such as standards for data quality and interoperability, digital archiving and long-term archival preservation, and for e-commerce transactions need to be factored in early, so that efficient and effective solutions can be developed in a timely fashion, rather than as an afterthought. Beyond formal standards and requirements, data centers need the ability to maintain strong links with both scientific and user communities, to develop close partnerships with their peer data centers, to look ahead strategically at future user needs and technological developments, and to control their own resources in ways consistent with their responsibilities.

Some of the higher-level functions are highlighted in the recent *ESDIS Data Center Best Practices and Benchmark Report* (Hunolt and Booth, 2001). For example, most of the 12 most commonly cited "best practices" derived from site visits and surveys of 15 "world class" data centers in the U.S. and abroad[3] deal with management and organizational issues, not specific service-level requirements, e.g.:

- *Have an established, formal technology refresh policy*.
- *Organize site by function, such that internal line organizations correspond to functional areas*.
- *Use vendor-independent, open-system architecture*.
- *Place scientists on-site who actively use the data.*
- *Place responsibility for site life-cycle acquisition, development, O&M with the operating site.*
- *Establish a process for management and science oversight of products and services.*
- *Emphasize data access simplicity and ease of use for users.*
- *Use formal service level agreements between a site and its customers.*

Similarly, the 1998 NRC DAAC review strongly recommended routine interactions with scientists (e.g., through a visiting scientist program and direct DAAC involvement in research activities), collaboration among data centers in a DAAC alliance, ongoing preparation and updating of strategic plans, and periodic external peer review.

These higher level functions can require significant time, resources, and effort that are not purely quantifiable in terms of data volume, throughput, and capacity and traditional views of sustaining engineering, center management, and user support. However, as evidenced by these and other reviews and studies, they are essential to the actual—and perceived—success of data centers in meeting user needs and meeting their responsibilities for sound data stewardship. We believe that

[3] Not including any DAACs.

the SEEDS initiative needs to take these high-level functions into account explicitly in its planning, design, and cost estimation efforts.

Specific Comments on the SEEDS Working Paper

1. Information technology (IT) security is only mentioned briefly. Given the intersection of the need to respond to continual challenges to IT security and the growing complexity of science data needs, the need to address IT security will evolve/expand over time. More attention should be given to these issues as part of data center operations and costing.
2. Long-term data preservation requirements should be taken into account at the start of data processing, not just at the end of the data life-cycle when the data are sent to the data center with the long-term archival responsibility.
3. There is no mention of billing and accounting/cost-recovery issues pertaining to data dissemination.
4. As data become more highly integrated, intellectual property rights will become more complicated. Current trends in data management point towards the addition of metadata pertaining to digital rights management as part of standard metadata.
5. The NewDISS concept paper outlines the need for flexible data access and integration, but there is little mention in the SEEDS working paper of data center requirements to support interoperability. One need is to participate actively in community processes to define and prototype core standards, which may in turn help determine appropriate levels of service.
6. The descriptions of the requirements are fairly static and do not address how data centers should address evolving expectations and standards.
7. The model underlying the working paper still largely reflects a fixed "product order and delivery" approach, not a more dynamic, flexible, Internet-based system driven by user needs to subset and integrate data from disparate sources on demand.
8. It would be clearer to present the LOS profile for each type of data center separately, rather than to let the reader try to compare the applicability matrix with the generic profile.
9. Science advisory needs are addressed, but there is no recognition of the growing range of consumers of ESE data. Certainly, the needs of educators will differ from those of journalists, policy-makers, and researchers. These types of end-users will need different types and levels of support and should be represented in external advisory groups.
10. Need a better definition of what counts as "off-site" for off-site backup storage. The definition of what is a "safe" distance may well be changing, as well as general expectations about what should and can be backed up and how frequently. Having real-time redundancy in processing systems will certainly become the norm in the business world and may well be appropriate for expensive satellite data systems.
11. The definition of active users is problematic for web-based delivery of data, or for a dynamic system driven by a flexible interface that allows users to subset data on-demand.

12. Maintaining system security, data integrity, and easy access may not always be mutually compatible.

References

Earth Science Enterprise (ESE). 2000a. *Exploring Our Home Planet: Earth Science Enterprise Strategic Plan.* Washington DC: National Aeronautics and Space Administration. Available online at: http://www.earth.nasa.gov/visions/stratplan/index.html.

Earth Science Enterprise (ESE). 2000b. *Understanding Earth System Change: NASA's Earth Science Enterprise Research Strategy for 2000-2010.* Washington DC: National Aeronautics and Space Administration. Available online at: http://www.earth.nasa.gov/visions/researchstrat/Research_Strategy.htm.

Hunolt, G., and A. Booth. 2001. *ESDIS Data Center Best Practices and Benchmark Report*. Greenbelt MD: Goddard Space Flight Center.

National Research Council (NRC). 1998. *Review of NASA's Distributed Active Archive Centers.* Washington DC: National Academy Press.

National Research Council (NRC). 2001. *The Science of Regional and Global Change: Putting Knowledge to Work.* Washington DC: National Academy Press.